LES FICHES CUISINE DE TONTON LULU

Nouveaux rebondissements dans l'affaire du "petit gros" :

LUTTE DES CLASSES DANS UN NUAGE ...

par Jacques LUBCZANSKI

J'ai eu l'occasion de vous entretenir d'un problème de fond :
"Qu'est-ce qu'un petit gros ?". Plus précisément :
Une "population" étant donnée par les poids et tailles des individus qui la composent, quel critère mathématique choisir pour déterminer l'ensemble des "petits gros" de cette population ?
D'ailleurs, le problème de fond était plutôt un problème de forme !
En effet, selon la forme qui résumait le mieux le "nuage" des points (*), on arrivait à telle ou telle conclusion.
Aujourd'hui, face au même problème, nous allons abandonner les idées de description géométrique du nuage.
Adieu donc l'art abstrait !
Place au vieux bon sens : "Qui se ressemble s'assemble !".
Et pour regrouper les individus d'une population en "classes" (dont l'une serait notre ensemble des "petits gros"), suivons ce vieux dicton, assemblons les points qui se ressemblent.
Le problème prend donc la tournure suivante : quel critère mathématique pour évaluer la ressemblance des individus, c'est-à-dire des points du nuage ?
Comment utiliser ce critère pour regrouper les points en "classes" ?
Et comment interpréter les résultats obtenus à la lumière du problème posé ?

On est au coeur de la branche des mathématiques qui s'appelle "l'Analyse des données". Mais rassurez-vous : il n'est pas question ici de tout dire sur ce sujet ! Faisons plutôt un petit bout de chemin, une promenade touristique dans ce domaine souvent -et à tort- méconnu.

(*) : le "nuage" est l'ensemble des points représentant, dans un plan cartésien, les individus de la population donnée : dans notre problème, chaque point a deux coordonnées (x,y) ; x est une mesure de la taille et y une mesure du poids.

Quel critère mathématique pour évaluer la ressemblance entre les points ?

- le critère le plus naturel : mesurer la distance entre deux points ; il nous mènera à une répartition de la population en "classes de distance" ;
- un autre critère possible : mesurer la distance, non plus entre les points, mais entre les classes d'une répartition : cela nous conduira à une hiérarchie entre classes, à une classification ;
- enfin, dernier critère évoqué dans cette fiche : mesurer la cohésion interne, la concentration de chaque classe d'une répartition, et la rendre optimale.

Comment utiliser le critère choisi pour regrouper les points en classe ?

Il s'agira ici avant tout de méthodes pratiques : dans tous les cas, la répartition de la population en classes sera le résultat d'un algorithme. En effet, si les méthodes sont toutes très simples dans leur principe, leur mise en oeuvre s'accompagne d'une quantité de calculs considérable, confiés bien entendu à un ordinateur : d'où la nécessité d'une résolution algorithmique. Il est même un cas où les ordinateurs actuels ne sont pas assez puissants, et où on doit se contenter pour l'instant d'une solution approchée !

Comment interpréter les résultats obtenus à la lumière du problème posé ?

La question de la pertinence des méthodes, du choix des différents paramètres est la plus délicate. C'est en conservant les mêmes données tout au long de cette fiche que nous pourrons faire les comparaisons nécessaires.

A NI DIEU NI MAITRE... : (Classes de distance)

a. L'outil : C'est la distance entre les points : deux points seront dits "voisins" si leur distance est inférieure ou égale à un seuil fixé $\bar{d}$.

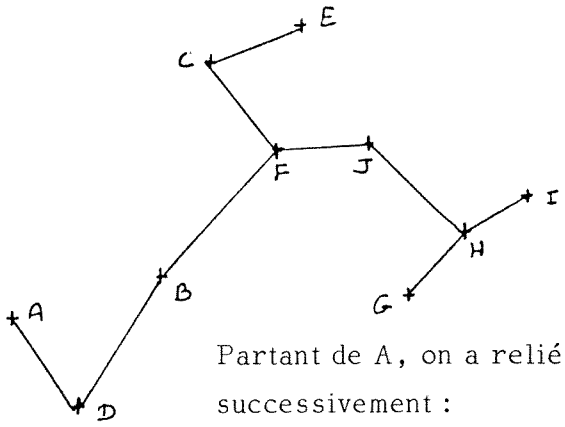
Les points voisins sont alors regroupés en "classes de distance" ; une "classe de distance $\bar{d}$" est simplement un ensemble (non vide) vérifiant la propriété : "si un point y est, alors tous ses voisins y sont aussi".

On obtient ainsi une répartition de la population en classes disjointes, de taille variable, non ordonnées : il s'agit d'amas anarchiques de points.

b. La mise en oeuvre : la détermination pratique des classes de distances se fait en deux temps :

. D'abord, on met en relation tous les points du nuage : on va dessiner un "arbre" dont ces points seront les noeuds. Pour que cet arbre soit le plus simple possible, on utilise l'algorithme suivant :

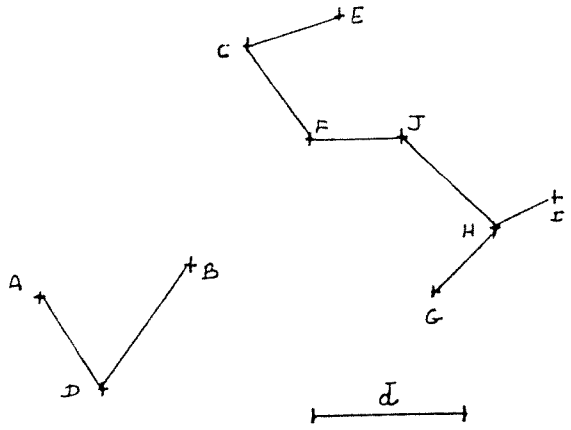
On part d'un point M_1, qu'on relie à son plus proche voisin M_2, d'où une "branche" M_1M_2. Alors M_3 est, parmi les points qui restent, le point le plus proche de la branche M_1M_2 : on le relie à cette branche par l'extrémité la plus proche, M_1 ou M_2.
D'où un (petit) arbre à deux branches :
M_4 sera le point le plus proche de cet arbre, parmi ceux qui restent, etc...



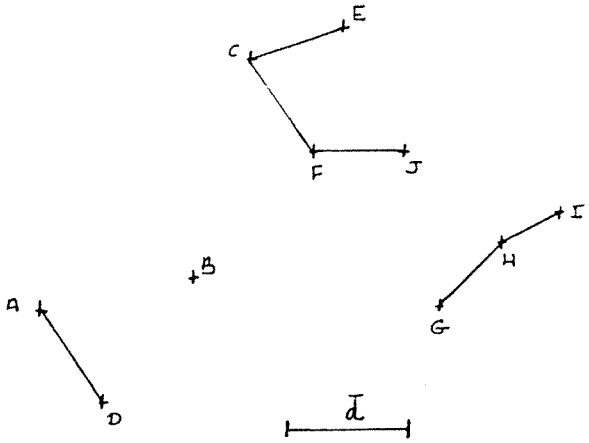
Partant de A, on a relié successivement :
D, B, F, J, C, E, H, I, G.

De proche en proche on construit ainsi un arbre reliant tous les points du nuage, et dont la longueur est minimale : c'est ce qu'on a fait pour les points du nuage ci-dessus, en partant du point A. (Mais en fait l'arbre obtenu ne dépend pas du point de départ.)

. Ensuite, ayant fixé le seuil $\bar{d}$, on coupe toutes les branches de longueur supérieure à $\bar{d}$: les "sous arbres" qui restent définissent les classes de distance cherchées.
Voici ce que cela donne pour notre nuage, avec deux choix différents de $\bar{d}$:



Classes de distance : $\{A, B, D\}$; $\{C, E, F, G, H, I\}$



Classes de distance : $\{A, D\}$; $\{B\}$; $\{C, E, F, J\}$; $\{G, H, I\}$.

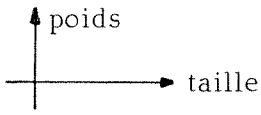
c . Pathologie des classes de distance : Notre exemple, pourtant très simple, suffit pour observer les deux maladies les plus courantes :

. l'anémie : les classes de très faible effectif, par rapport à celui de la population, ne sont pas très significatives (par exemple ici $\{B\}$) ; leurs points sont en fait des isolés.

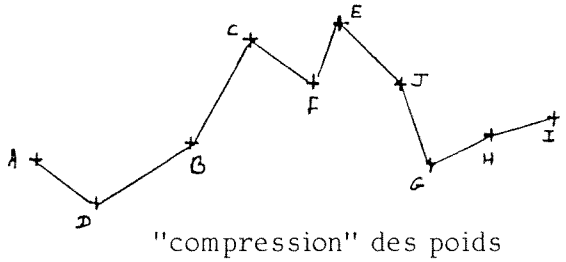
. La filiformie : dans une même classe, on peut trouver des points très éloignés les uns des autres, mais reliés par une chaîne de points voisins : dans notre exemple de gauche, E et G sont dans la même classe, mais pas F et B !

Il existe un remède à ces deux maladies : c'est d'éliminer les isolés, c'est-à-dire les points qui n'ont pas assez de voisins ; après avoir fixé le seuil $\overline{d}$, on fixe un seuil n : un point qui a moins de n voisins sera déclaré "isolé" et impitoyablement supprimé, avant même de tracer l'arbre : on gagne ainsi en temps de calcul et en "interprétabilité". Les effectifs des classes sont alors tous supérieurs à n ; dès que $n \geqslant 3$, les classes filiformes disparaissent.
En pratique, c'est donc un choix judicieux des deux seuils $\overline{d}$ et n qui assure une répartition satisfaisante : plusieurs essais peuvent être nécessaires (et en tous cas un calcul préalable du nombre de voisins de chaque point pour différents seuils $\overline{d}$).

d. Interprétation des résultats : Dans notre exemple, c'est certainement le choix pour $\overline{d}$ de l'exemple de droite, et pour n de la valeur 1 qui donne une répartition correcte : $\{A,D\}$; $\{C,E,F,J\}$; $\{G,H,I\}$ .B a été éliminé car il n'a pas de voisins. Si on se souvient que la taille est en abscisse et le poids en ordonnée, que peut-on dire ? $\{C,E,F,J\}$ est le groupe des "lourds". $\{G,H,I\}$ est le groupe des "grands". Quant à $\{A,D\}$ ce sont, par rapport aux autres, les "petits légers". Et les trois groupes sont bien séparés.
Où sont passés les petits gros ?

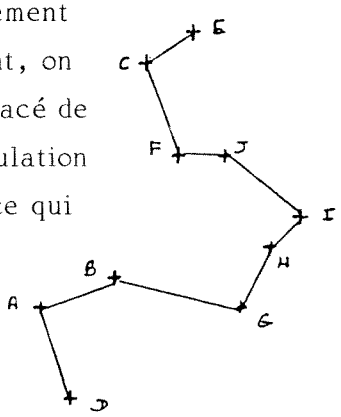


e. Attention aux échelles : Il faut aussi se souvenir qu'on a choisi des échelles de représentation, aussi bien pour le poids que pour la taille.
Un changement d'échelle, sur l'une ou l'autre des coordonnées, modifiera les distances entre les points et la répartition obtenue :



"compression" des poids dans un rapport 2.

Mais si le changement est trop important, on trouve, dès le tracé de l'arbre, une population filiforme, c'est ce qui se passe ici.



"compression" des tailles dans un rapport 2.

Il est rassurant de vérifier que des "compressions" intermédiaires entre les deux exemples extrêmes dessinés, ne donnent en fait qu'une seule répartition en classes "équilibrées", qui n'est autre que celle trouvée plus haut : $\{A,D\}$; $\{E,F,C,J\}$; $\{G,H,I\}$.
C'est peut-être une manifestation de l'équivalence des différentes métriques du plan !

B ORDRE ET DISCIPLINE... : (Distance de classes)

Le but avoué de cette méthode est d'établir une classification hiérarchique des points du nuage : la population sera divisée en classes ; les classes en sous-classes ; les sous classes en "sous-sous classes" etc... jusqu'aux classes réduites aux individus. Un exemple célèbre du genre est la classification des zoologistes (vertébrés, invertébrés...).

a. L'outil : C'est la distance, non plus entre les points mais entre les classes : si $\mathcal{C}_1$ et $\mathcal{C}_2$ sont deux classes, la distance $d(\mathcal{C}_1, \mathcal{C}_2)$ est la plus petite distance possible entre un point de $\mathcal{C}_1$ et un point de $\mathcal{C}_2$: ce qu'on peut écrire :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \inf_{\substack{x_1 \in \mathcal{C}_1 \\ x_2 \in \mathcal{C}_2}} d(x_1, x_2)$$

b. La mise en oeuvre : C'est un algorithme : on part de classes réduites à un point et à chaque étape, on fusionne deux classes (ou plus) : celles qui sont les plus rapprochées : d'où une nouvelle répartition.
Voyons cela sur notre exemple :
Il faut d'abord dresser le tableau des distances entre les points : en pratique ici, on a inscrit les carrés des distances, pour conserver des chiffres simples (des nombres entiers), ce qui ne modifie pas la classification.
Si l'effectif de la population est n, il faut donc calculer $\frac{n(n-1)}{2}$ distances.
Et c'est tout.
En effet, les distances de classes qui vont nous servir sont des nombres de ce tableau.

	A	B	C	D	E	F	G	H	I	J
A	0	26	113	13	181	106	169	229	298	169
B		0	53	25	39	32	65	101	148	77
C			0	146	10	13	100	100	125	34
D				0	208	113	130	194	261	164
E					0	17	90	74	85	20
F						0	41	45	68	9
G							0	8	25	26
H								0	5	18
I									0	29

1ère étape : on fusionne H et I en HI, car d(H, I) = 5 : c'est le plus petit nombre du tableau. Pour continuer, il suffit de récrire le tableau des distances de classe, c'est-à-dire de remplacer, dans le tableau précédent, les lignes H et I par une ligne HI et les colonnes H et I par une colonne HI, en prenant le plus petit des deux nombres : le reste du tableau ne change pas. (Ici la colonne HI est en fait celle de H, car I est plus éloigné que H de tous les autres points).

	B	C	D	E	F	G	HI	J
A						169	229	
B						65	101	
C						100	100	
D						130	194	
E						90	74	
F						41	45	
G						0	8	
HI							0	18

1ère étape

2ème étape : on cherche le plus petit nombre du tableau : c'est d(G,HI) . Si on fusionne G et HI en GHI, le tableau rétrécit encore d'une ligne et d'une colonne

	B	C	D	E	F	GHI	J
A						169	
B						65	
C						100	
D						130	
E						74	
F						41	
GHI						0	18

2ème étape

3ème étape : le plus petit nombre du tableau est 9 : on fusionne F et J

4ème étape : le plus petit nombre du tableau est 10 : on fusionne C et E

5ème étape : le plus petit nombre du tableau est 13, qui figure deux fois : on fusionne A et D d'une part, C et F d'autre part.

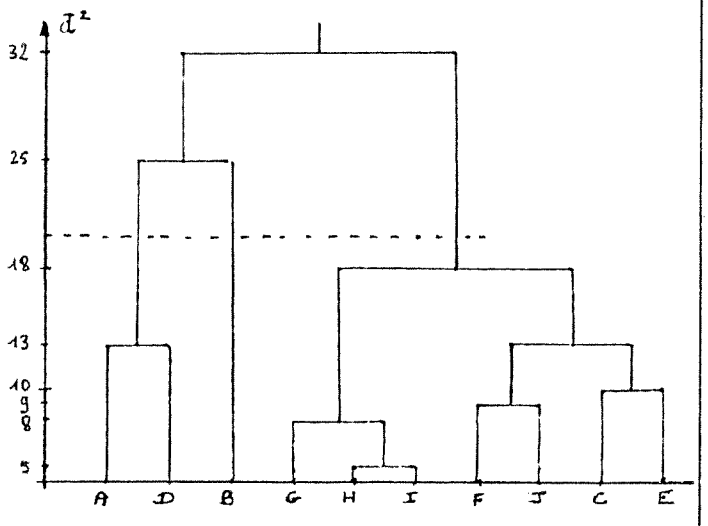
En pratique, il est inutile de récrire les tableaux à chaque étape : il suffit de suivre les nombres du tableau, du plus petit au plus grand : après 13 vient 17, distance entre F et E : mais F et E sont déjà dans la même classe : au suivant !

C'est 18 = d(J,H) : on fusionne GHI et FJCE.

Puis 25 = d(B,D) et aussi d(G,I) : G et I sont déjà ensemble ; on fusionne B et AD. Tiens il ne reste plus que deux classes ABD et GHIFJCE : on les fusionne !

Pour résumer, on dresse un arbre, que les savants appellent dandogramme : ce dessin représente la classification de la population étudiée, en classes et sous classes...

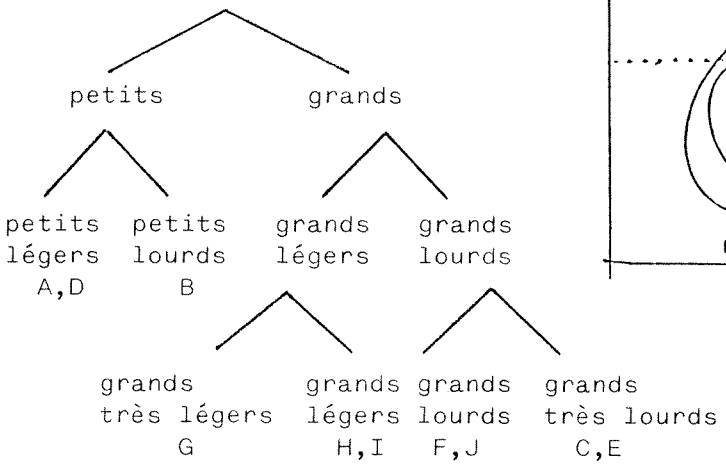
En coupant l'arbre à une hauteur quelconque, il tombe des branches ! On obtient la répartition en classes correspondant à cette hauteur, à ce seuil :

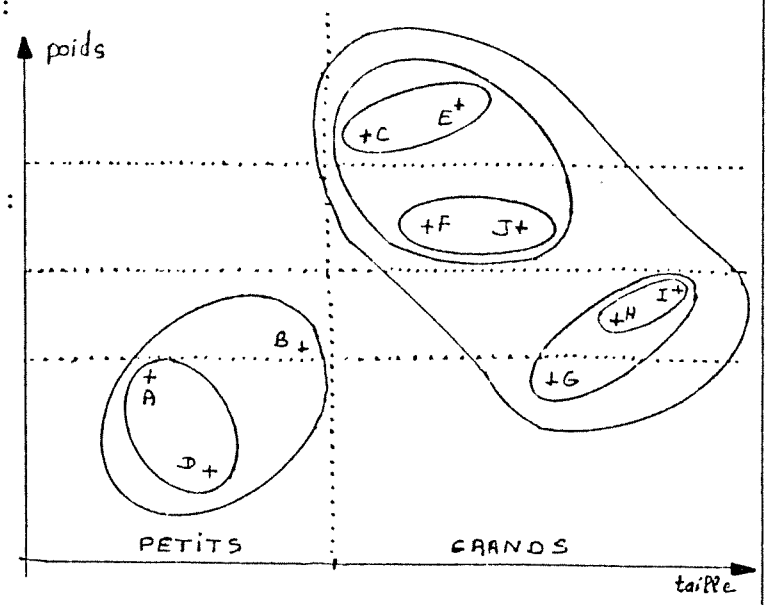


Par exemple, si on scie à la hauteur 20, il tombe trois branches :
AD ; B ; GHIFJCE.

c. Interprétation des résultats : on peut représenter sur le nuage la classification obtenue :

Les lignes de niveau de la taille (droites verticales) et du poids (droites horizontales) permettent alors d'interpréter chaque "fourche" du dendogramme :





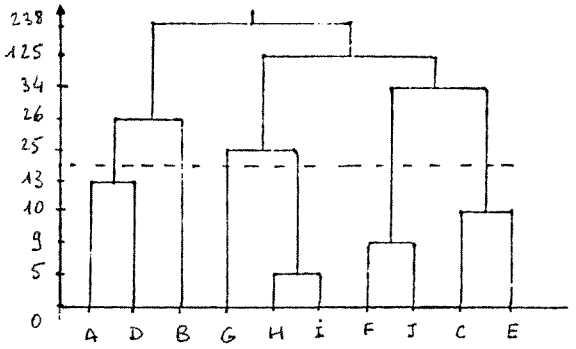
Et si on cherche les "petits gros" parmi les petits lourds, un seul candidat : B.
De même si on cherche les grands maigres parmi les grands très légers, G est un grand maigre.

d. Pathologie de la distance de classes :
- on en arrive à opposer B, petit gros, à G, grand maigre, alors que G est plus proche de B que de C ;
- on aurait pu séparer AD et B, (ainsi que FJCE et GHI) par la taille : on aurait obtenu des "très petits" (AD) et des "petits" (B) : en cherchant les "petits gros", c'est A qu'on aurait choisi, comme le plus lourd des "très petits".
On peut attribuer ces défauts de la méthode au choix qu'on a fait de la distance entre classes, qui favorise la fusion de deux classes dès qu'elles possèdent des points proches : le risque est alors de trouver dans une même classe des points très éloignés.

e. Remèdes homéopathiques : il y a deux autres façons classiques de mesurer la distance entre classes, qui évitent la "pathologie" rencontrée :

- d ($\mathcal{C}_1$, $\mathcal{C}_2$) peut être la plus grande distance possible entre un point de $\mathcal{C}_1$ et un point de $\mathcal{C}_2$:

$$d(\mathcal{C}_1, \mathcal{C}_2) = \sup_{\substack{x_1 \in \mathcal{C}_1 \\ x_2 \in \mathcal{C}_2}} d(x_1, x_2)$$

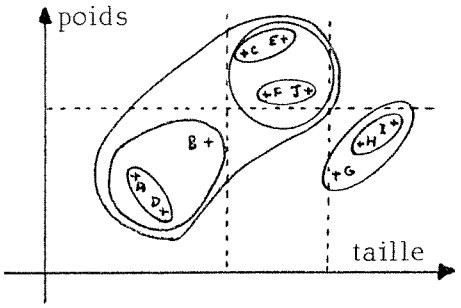


Avec cette distance, les classes ne peuvent fusionner que si tous leurs points sont proches. Dans notre exemple, on arrive alors à une hiérarchie différente, mais à la même classification ! Etonnant non ?

- d ($\mathcal{C}_1$, $\mathcal{C}_2$) peut enfin être simplement la moyenne des distances des points :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{n_1 \times n_2} \sum_{x_1 \in \mathcal{C}_1} \sum_{x_2 \in \mathcal{C}_2} d(x_1, x_2)$$

où n_1 et n_2 sont les effectifs de $\mathcal{C}_1$ et $\mathcal{C}_2$.



On obtient bien sûr une hiérarchie et une classification différentes des précédentes, où on ne peut plus caractériser nettement des "petits lourds". GHI sont des grands maigres : ABD est plus petit et plus léger que FJCE. Il n'y a plus de petits gros !

f. Dernier examen clinique : quels résultats résistent à la valse des distances ?

C'est la répartition en trois classes ADB, GHI et FJCE : en effet ces classes ne fusionnent que pour des valeurs élevées de $\bar{d}$: avec la distance moyenne, il faut attendre $d^2 = 94$, alors que ces classes sont constituées dès que $d^2 = 21$; et il en est de même avec les autres distances. On retrouve -mais est-ce étonnant ? -une des répartitions obtenues avec la méthode du paragraphe précédent. Avec une hiérarchie (variable) en prime !

C QUI M'AIME ME SUIVE... (Critère de variance interne)

Une autre façon de juger si un ensemble de points, une classe, forme un "amas" est de prêter attention à sa cohésion interne : les points sont-ils rassemblés autour d'un "centre" ou au contraire sont-ils dispersés ?

L'outil : c'est la notion classique de variance, que nous allons retrouver : en effet, on va mesurer la cohésion d'un ensemble de points en calculant d'abord le centre de

cet ensemble, puis la distance moyenne à ce centre : plus celle-ci sera faible, plus les points seront regroupés autour du centre.
Et si la moyenne des distances est calculée de façon non pas arithmétique mais quadratique (c'est-à-dire la racine carrée de la moyenne des carrés), on retrouve la notion de variance totale d'un nuage (voir "Qu'est-ce qu'un petit gros" § 1 & 2). Si $\mathcal{C}$ désigne la classe étudiée et G son centre, la variance de $\mathcal{C}$ est $V = \frac{1}{n} \sum_{M \in \mathcal{C}} d^2(M,G)$, où n est l'effectif de $\mathcal{C}$.

Si à présent, on considère toute la population qu'on veut répartir en classes le plus cohérentes possibles, on peut, pour chaque répartition, calculer la somme des variances de chaque classe, et chercher à rendre cette somme la plus petite possible.
Cette somme porte parfois le nom de variance intraclasses, par opposition à la variance interclasses, qui est la variance de l'ensemble des centres des classes. En anglais, les termes sont plus clairs : on parle de variance "within" et de variance "between". Pour ma part, je dirai simplement variance interne (pour mesurer la cohésion à l'intérieur de chaque classe) et variance externe (pour mesurer la cohésion entre classes).
Or, il se trouve que si V est la variance totale de la population, $V = v_i + v_e$ où v_i et v_e sont les variances interne et externe de la répartition.
Autrement dit, on pourra soit minimiser v_i soit maximiser v_e.

La méthode théorique : elle est simple : l'effectif de la population étant fini, il suffit de faire calculer à l'ordinateur la variance v_i pour toutes les répartitions possibles : celles-ci sont en nombre fini donc il y en a au moins une qui minimise v_i : c'est la meilleure répartition en amas.
Malheureusement, il y a deux écueils :
- la répartition minimisant v_i est connue : c'est la répartition... en singletons, pour laquelle $v_i = 0$; et plus généralement : plus il y a d'amas, plus v_i est petite : notre recherche de minimum doit donc se faire à nombre d'amas constant : on se fixe d'avance k nombre de classes de la répartition et c'est parmi les répartitions en k classes qu'on cherche celle qui minimise v_i.

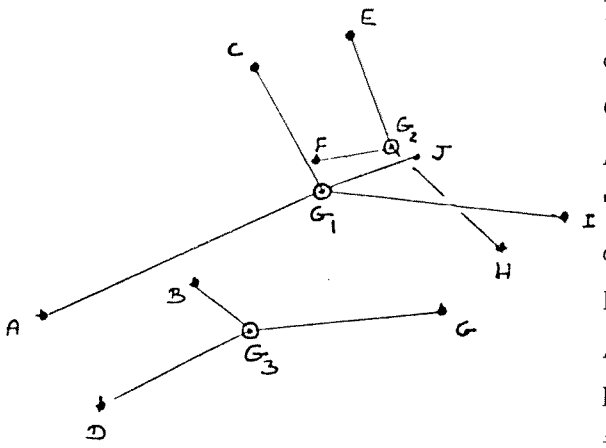
- seulement voilà : même en fixant le nombre k de classes, il y a une quantité énorme de répartitions possibles.
Tenez, dans notre exemple, il y a 34105 répartitions des 10 individus en 4 classes.
La formule générale est $S_m(k) = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} C_k^j \, j^m$ où m est l'effectif de la population et k le nombre de classes.

Pour nous donner une idée, $S_{15}(5) \simeq 2{,}1 \times 10^{8}$

$S_{25}(5) \simeq 2{,}4 \times 10^{15}$

Ce nombre croît effroyablement vite avec m : alors si on étudie une population d'effectif un peu sérieux, on dépasse très vite les capacités des ordinateurs actuels !

La mise en oeuvre : faute de pouvoir calculer la répartition "optimale", on va chercher une répartition "suboptimale" : en d'autres termes, on va utiliser un algorithme qui, partant d'une répartition donnée, la modifie à chaque étape de façon à diminuer la variance interne v_i : décrivons cet algorithme pour notre exemple :



Partons d'une répartition quelconque, en trois classes ACJI, BDG et EFH.

Calculons les trois centres G_1, G_2 et G_3.

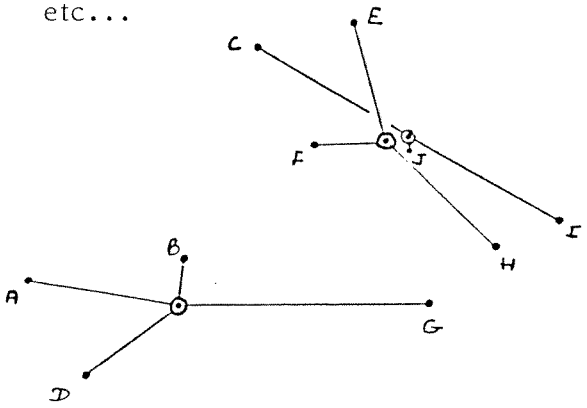
A est plus près de G_3 que de son centre G_1: "qui m'aime me suive !" lui crie G_3, et voici A qui change de classe...

La nouvelle répartition est donc :

ABDG, CJI et EFH.

Les centres ne sont plus les mêmes et la variance interne a diminué.

On recommence : on prend un par un les points nuage ; dès que l'un d'entre eux est plus près d'un autre que du sien, on le change de classe : on recalcule les nouveaux centres, etc...

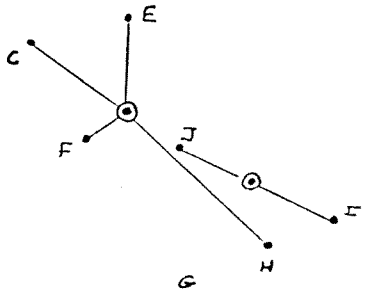


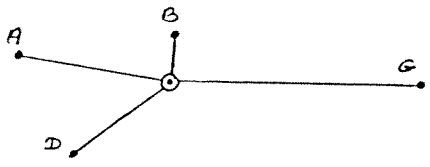
Voici les étapes de l'algorithme pour notre exemple :

B ne change pas de classe,

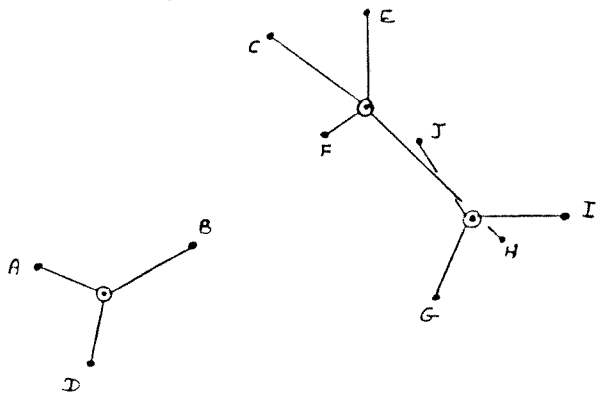
C rejoint la classe EFH...

D'où la configuration :

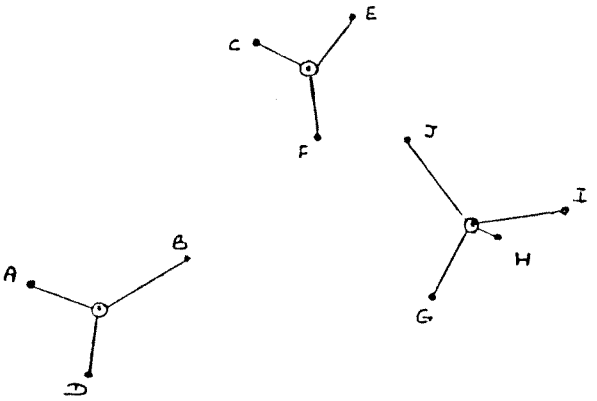




Puis G rejoint JI :



Enfin, H change à son tour de classe et on arrive à cette répartition, que l'algorithme ne modifie plus :

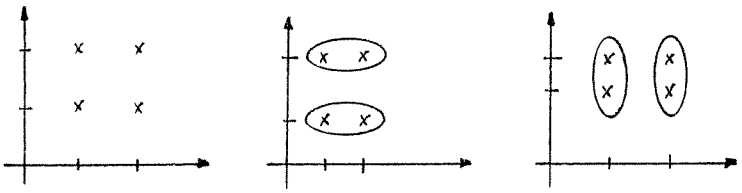


(on n'était d'ailleurs pas obligé de suivre l'ordre alphabétique... et le résultat serait différent avec un autre ordre !)

Pathologie des centres "mobiles" :

- Sensibilité aux modifications d'échelles : il faut faire attention à ne pas créer artificiellement des amas par un choix maladroit des échelles de représentation :

En effet, que penser de ces trois représentations des mêmes données avec des échelles différentes ?



- Comment choisir d'avance le nombre k d'amas ? Ce ne sont pas les chiffres qui vont répondre (le minimum de v_1 décroît quand k augmente) mais la pertinence de la répartition obtenue par rapport aux questions étudiées.

- Enfin, la solution obtenue est suboptimale et non optimale : elle dépend de la répartition initiale choisie : il faut si possible faire plusieurs essais pour choisir ensuite la solution la plus interprétable.

Interprétation des résultats :

D'autres essais, à partir de répartitions initiales différentes, donnent des répartitions finales différentes ; celle qu'on a obtenue à la page précédente ne donne pas la variance interne minimale, réalisée pour la répartition ADB, CEFJ et GHI ; on a déjà rencontré et interprété celle-ci : ne nous y attardons pas.

Un autre choix de k donne des classes malingres (doubletons pour k = 4, singletons pour des valeurs supérieures) ou au contraire des classes opulentes (k = 2 ... et k = 1). Le choix de k = 3 semble donc pertinent.

Le retour de la hiérarchie :

Le critère de variance interne, qu'on a utilisé, peut aussi servir à la méthode de classification hiérarchique vue plus haut : en effet on peut décider à chaque étape de l'algorithme de classification de fusionner les deux classes de façon à "gagner" le plus possible de variance interne (ou à "perdre" le moins possible de variance externe).
Cela revient à mesurer la distance entre classes par cette variation de variance :
si $\mathcal{C}_1$ et $\mathcal{C}_2$ sont deux classes d'effectifs n_1 et n_2, de centre G_1 et G_2, on montre que $\Delta v_i = \frac{n_1 n_2}{n_1+n_2} d^2(G_1,G_2)$ lors de la fusion de ces deux classes (et $\Delta v_e = -\Delta v_i$).
Alors on prend $d(\mathcal{C}_1, \mathcal{C}_2) = \frac{n_1 n_2}{n_1+n_2} d^2(G_1,G_2)$ et l'algorithme s'applique.

D CONCLUSION

Voilà. La promenade au pays de l'analyse de données touche à sa fin.

Certains dirons peut-être que c'était se donner beaucoup de mal pour retrouver des amas qu'on distingue à l'oeil nu sur le nuage.
Je répondrai que les méthodes présentées ont l'avantage de rester valables en dimension 3, 4 ... n, ce qui arrive souvent dans la pratique, mais où l'oeil n'est d'aucun secours.
Et si j'ai voulu les présenter en dimension 2, c'est pour montrer combien, pour finir, ces méthodes sont simples dans leur principe, et efficaces dans leur action.

Et surtout, cette partie des mathématiques est sans doute la plus utilisée par les non-mathématiciens, pour confirmer ou infirmer des hypothèses, pour affirmer ou nier des théories : cela peut être lourd de conséquences, quand parfois, les mathématiques servent d'alibi à l'esprit critique.

Je pense que cela mérite bien un détour.